Financial Economics Research; Vol.3, No.4; 2026
E-ISSN: 3006-3663 P-ISSN: 3006-3655
DOI: https://doi.org/10.70267/fer.v3n4.8496
Published by: Zeus Press

Cross-Sectional Return Prediction in China's A-Share Market Based on Lasso and XGBoost: A Comparison with the Fama–MacBeth Multi-Factor Model

Xichun Zhang*

School of Mathematics and Statistics, North China University of Water Resources and Electric Power, Zhengzhou 450046, China

**Corresponding author: Xichun Zhang.*

Abstract

In response to the problem of the widespread use of mixed prediction performance metrics in machine learning stock selection literature, the lack of empirical tests on the applicability boundaries of linear and tree-based models under low-dimensional factor settings, and the disconnection of most high-dimensional factor empirical conclusions from the actual scenarios of small and medium-sized quantitative institutions, this paper selects the monthly panel data of A-shares from 2013 to 2023 to avoid the leakage of look-ahead information, and builds a unified rolling-window forecasting framework of FM, Lasso-FM, and XGBoost. The model hyperparameters are selected based on the existing literature to ensure a fair comparison among the three models. It relies on OOS-R^2, IC, Rank-IC and long-short portfolio Sharpe ratio to carry out multi-dimensional evaluation, and carries out analysis in combination with the heterogeneity of market value grouping and SHAP decomposition. In addition, multiple sets of robustness checks are conducted. The empirical evidence shows that under the sample setting of the five basic factors in this article, the mean out-of-sample OOS- R^2 values of all three models are negative, the absolute return prediction performance is weak, and there is no significant statistical superiority or inferiority between the models; However, there is a deviation in the evaluation index, and XGBoost performs better in cross-sectional stock ranking and long–short portfolio performance, which is rooted in the fact that OOS-R^2 focuses on absolute error, and IC only focuses on the relative return strength of stocks; Based on the observation of this sample, it is difficult for XGBoost to release nonlinear modeling capability, and greater model flexibility has the possibility of inducing out-of-sample overfitting. The forecast stability of large-cap stocks is significantly higher than that of small-cap stocks, and the market value and ROA are important prediction features of the model. This paper explains the intrinsic mechanism of multi-indicator deviation, defines the applicable boundaries of factor dimensionality of different models, and provides a reference for target lightweight modeling for small and medium-sized institutions.

Keywords

cross-sectional return prediction, Fama–MacBeth, Lasso, XGBoost, SHAP, out-of-sample OOS- R^2

1. Introduction

Machine learning methods are widely used in the cross-sectional return prediction research of A-share stocks, providing new analytical tools for quantitative stock selection. Extreme Gradient Boosting (XGBoost) shows a good predictive effect in a large number of high-dimensional empirical studies with the ability to mine the high-order interaction of factors [1,2,3]. Most of the existing domestic related studies are based on the hundred-dimensional derived factors to carry out backtesting to verify the out-of-sample prediction advantages of XGBoost. However, the prerequisites of this kind of research deviate from the actual modeling scenario of small and medium-sized institutions in China. Small and medium-sized institutions can often only obtain a small number of basic factors such as financial and market variables, and it is difficult for the research conclusions obtained in the high-dimensional environment to be directly transferred to low-dimensional modeling practice [4]. At the same time, some literature often mixes many indicators such as out-of-sample R^2, OOS-R^2, Information Coefficient (IC), and rank IC in the evaluation of the model performance. The indicator divergence phenomenon of ”out-of-sample absolute fitting fails, but the stock portfolio can still obtain positive returns“ has not been systematically explained; Most studies are based on high-dimensional data sets to derive the judgment that the nonlinear model is comprehensively superior, but rarely test whether the conclusion is valid under the condition of low-dimensional basic factors, and lack the normative empirical test of the applicability boundaries of linear and tree-based models [5]. In addition, some similar studies also have empirical problems such as look-ahead information leakage, inconsistence of multi-model comparison benchmarks, and confusion between the internal prediction weight of the model and the factor risk premium.

In view of the above research shortcomings, this article aims to clarify the inherent logical differences of different prediction and evaluation indicators, explore the relative performance of linear and machine learning models under the constraints of low-dimensional basic factors, and supplement the empirical experience of A-share low-dimensional scenarios. This article selects the monthly panel data of A-shares from 2013-2023, uses the design of t-1 period public factor delay to predict returns at time t+1 to avoid look-ahead bias, builds three types of prediction models of Fama-MacBeth (Fama-MacBeth, FM), Least Absolute Shrinkage and Selection Operator (Lasso-FM), and XGBoost, and builds a unified rolling-window backtesting framework. The model hyperparameters are selected based on the existing literature to select fixed values to ensure the fairness of the comparison of the three types of models. In combination with SHapley Additive exPlanations (SHAP), we carry out factor contribution analysis, and at the same time, we combine the circulation market value grouping for heterogeneity analysis, and set up multiple groups of independent robust tests to improve the reliability of the results, in order to provide reference ideas for lightweight modeling for small and medium-sized institutions with limited computing power and factor resources. It should be noted that the empirical conclusion of this article only adapts to the sample and factor settings of this study, and is not universal.

2. Literature Review and Theoretical Framework

2.1 Literature Review

The classical linear analysis framework of asset pricing includes CAPM, Fama-French factor model, and Fama-MacBeth cross-section regression. Previous international studies have shown that under the massive supply of features, XGBoost relies on high-order nonlinear interaction to improve the ability to predict out-of-sample returns, but the gain is highly dependent on sufficient factor input [6]. There is an obvious division in the relevant domestic research. Wu Shinong and Xu Nianxing believe that the speculative behavior of A-share retail investors raises the monthly return noise, and the high-degree-of-freedom nonlinear model is easy to fit random fluctuations with no economic meaning [1]. Li Bin verified the data noise reduction effect of L1 regularization, but did not make a parallel comparison with XGBoost [4]. Wang Xi and others only test the low-dimensional Lasso model, and lack the machine learning reference benchmark [7]. Gu Songnan, Qiao Zheng and others rely on the high-dimensional derivative factor to draw the superior conclusion of XGBoost, which is difficult to adapt to the low-factor practical scenarios of small and medium-sized institutions [2,3]. A domestic review has pointed out that machine learning can improve multi-factor pricing in a big data environment, but the comparative empirical evidence of linear and tree models under low-dimensional basic factor conditions is still insufficient [5]; The existing autoencoder-based machine learning asset pricing research points out that the strong fitting ability of machine learning is easier to capture market noise, and the model’s predictive performance is limited under low-dimensional conditions [8]. Herd and noise trading are important features of the A-share market, which will amplify the risk of model overfitting [9]; the empirical evidence of deep learning and tree models in the domestic market is mostly based on high-dimensional feature data sets [10,11]. The existing literature lacks parallel comparison of low-dimensional basic factors under a unified framework, and the interpretation of the economic transmission mechanism of empirical results is insufficient, which is supplemented by this article.

2.2 Core Theoretical Mechanisms

2.2.1 Fundamental Logic of Factor-Based Prediction

The five mature forecast indicators of log market capitalization LnSize, book market value ratio BM, return on assets (ROA), monthly turnover rate and monthly momentum are selected, which correspond to the five return-generating mechanisms of liquidity premium, valuation repair, sustainable profit, short-term speculation and price trend respectively. Extensive evidence from China has demonstrated the effectiveness of these factors in monthly return prediction.

2.2.2 Applicability Boundaries of Linear and Tree-Based Models

FM and Lasso-FM only capture factor linear association; The core value of XGBoost is to mine multi-factor high-order interactions, and this ability is only fully released in the hundred-dimensional high-factor pool. Under the sample setting of the five basic factors in this article, there are very few interactive relationships that can be mined, and the inherent advantages of the tree model are difficult to give full play. At the same time, the monthly return noise of A-shares is high, the XGBoost model flexibility is higher, and is more likely to memorize random noise in the training data, leading to overfitting and weaker out-of-sample predictive performance.

Traditional Lasso is mostly used to screen redundant variables in high-dimensional data sets. There are no large number of redundant features under low-dimensional conditions in this article. The L1 regularization no longer assumes the screening function. It only slightly alleviates the mild multicollinearity of LnSize and BM, compresses the low contribution coefficient of noise-driven, and improves the stability of model generalization. The full text weakens the relevant strong expression of “Lasso high-dimensional screening, XGBoost complex nonlinear mining”, which fits the five-factor sample constraints.

2.2.3 Potential Channels of Heterogeneous Predictive Performance across Market Capitalization Groups

Large-market individual stock institutions hold positions, analyst tracking coverage is sufficient, information disclosure is standardized, and fundamental forecast signals are less interfered by short-term speculation; Small-sized markets are mainly retail transactions, and the subject speculation is frequent. Superimposed information disclosure is lagging behind, institutional coverage is insufficient, forming information asymmetry, and effective income signals are easily diluted by random capital fluctuations. This article does not introduce institutional shareholding and analyst coverage agent variables to carry out intermediary tests. The stratification of large and small plate indicators is only for correlation observation, and information asymmetry is only a potential explanation path, without causal judgment. The relevant intermediary conduction is left for further expansion.

2.2.4 Mechanism of Rolling Window Constraints

The benchmark rolling window is set to 60 months, taking into account the sample capacity and data timeliness; the 36- and 48-month windows are only used for model hyperparameter sensitivity testing, and are not standardized independent robust testing.

3. Research Design

3.1 Data and Variable Definitions

The sample period covers monthly panel data for A-shares from CSMAR, spanning January 2013 to December 2023. Data preprocessing rules: Exclude ST, *ST, and delisting-watch stocks; remove samples with missing data for the factor at period t or the return at period t+1. Individual stock excess returns are calculated using the 3-month time deposit interest rate; extreme values for variables such as return, turnover rate, and momentum are truncated, and all continuous variables undergo two-sided winsorization (at the 1% and 99% levels) to eliminate extreme outliers.

a) Dependent variable: $Mret_{i,t+1}$: The raw monthly return of stock *i* for month *t*+1. Throughout the process, factors from period *t* are used with a one-period lag to predict returns for period *t*+1; using contemporaneous factors to explain contemporaneous returns is prohibited to avoid look-ahead bias and information leakage.

b) Predictors (all based on month-end observations for period *t*): Log of market capitalization (LnSize), book-to-market ratio (BM), return on assets (ROA), monthly turnover rate, and monthly momentum.

c) Grouping variable: Month-end tradable market capitalization of individual stocks; stocks are classified into large-cap and small-cap groups based on the monthly sample median.

3.2 Model Specification

The model inputs, prediction targets, and time-series partitioning are completely consistent across all three types of models, ensuring a fair comparison.

3.2.1 Traditional Fama–MacBeth Cross-Sectional Regression Model

Cross-sectional linear regressions of individual stocks are estimated monthly, and the resulting time-varying coefficients are used to generate return forecasts for the following month:

$$Mret_{i,t+1} = \alpha_t + \sum_{j=1}^{J} \beta_{j,t} \, Factor_{i,j,t} + \varepsilon_{i,t+1} \tag{1}$$

Where: $Mret_{i,t+1}$ is the raw monthly return of stock *i* for month *t*+1; α_t is the monthly intercept term; $\beta_{j,t}$ is the cross-sectional regression coefficient for the *j*-th factor in period *t*; $Factor_{i,j,t}$ is the value of the *j*-th factor for stock *i* in period *t*; $\varepsilon_{i,t+1}$ is the random disturbance term; and $\sum$ is the summation operator.

3.2.2 Lasso–Fama–MacBeth Linear Shrinkage Model

Introducing L1 regularization into the linear least-squares loss function shrinks the coefficients of low-contribution factors and mitigates multicollinearity:

$$\min_{\beta} \sum_{i=1}^{N} \left(Mret_{i,t+1} - \alpha_t - \sum_{j=1}^{J} \beta_{j,t} \, Factor_{i,j,t}\right)^2 + \lambda \sum_{j=1}^{J} |\beta_{j,t}| \# \tag{2}$$

In the equation, *min* denotes the minimization operation; λ is the L1 regularization penalty coefficient; and the other symbols have the same meanings as in Equation (1). This paper selects a fixed penalty coefficient for modeling, drawing upon existing quantitative research on A-shares.

3.2.3 XGBoost Nonlinear Ensemble Prediction Model

Using five factors from period t as inputs, multiple decision trees are iteratively fitted to model the return for period *t*+1.

$$y_{i,t+1} = \sum_{k=1}^{K} f_k \left(Factor_{i,t}\right) \tag{3}$$

Where $y_{i,t+1}$ is the predicted return of stock *i* for month *t*+1; $f_k(\cdot)$ is the *k*-th decision tree model; *K* is the total number of decision trees; $\sum$ is the summation symbol. This paper sets the model's fixed hyperparameters based on similar literature; under low-dimensional constraints, only simple factor interactions can be captured.

3.3 Standardized Rolling-Window Procedure

This paper employs a rolling window approach for out-of-sample forecasting. The benchmark regression utilizes a fixed 60-month training window that rolls forward month by month. For each period, the model is trained using the full history of monthly samples within the rolling window to forecast individual stock returns for the subsequent month; this strictly adheres to the principle of using factors from period t to predict returns for period t+1, thereby avoiding look-ahead bias. The hyperparameters for Lasso and XGBoost were selected based on fixed values from domestic A-share return prediction literature; no independent validation set was used for grid search. To examine the sensitivity of the conclusions to the window length, the robustness check includes two additional sets of rolling windows—48 months and 36 months—using which the aforementioned backtesting process is repeated.

Fixed model parameters: Lasso penalty coefficient α = 0.0001; XGBoost number of trees n_estimators = 100, maximum tree depth max_depth = 3. Ordinary Least Squares (OLS) estimation is used for the Fama-MacBeth procedure.

3.4 Multidimensional Evaluation Framework for Predictive Performance

This article adopts four-layer evaluation standards and discards the single OOS-R^2 indicator: (1) Out-of-sample fitting: monthly OOS-R^2; (2) Sorting index: information coefficient IC, Rank Information Coefficient Rank IC; (3) Portfolio performance: decile long–short portfolio returns, annualized Sharpe ratio, deduct 0.3% bilateral cost income; (4) Statistical test: A Newey–West adjusted *t*-test is conducted on the monthly differences in the performance metrics to determine whether the differences between models are statistically significant.

3.5 Distinction between the Validation Frameworks

Distinguish between window sensitivity and independent robustness test to avoid conceptual confusion: window adjustment only verifies the tolerance of the model to parameters, and does not change the core settings of samples and factors; The five types of robust tests change the research settings from the multi-dimensionality of sample, return definition, factor construction, benchmark model and transaction cost, which is used to prove the universality of the conclusion, and the two tests are logically independent.

a) Window sensitivity test: 36, 48-month rolling length, only for parameter endurance verification, not a robust test.

b) Five types of independent robustness tests: segmented sub-sample, excess return substitution, factor industry neutrality, gradual regression benchmark, and multi-stage transaction cost.

Expansion analysis: the important prediction features of market value grouping heterogeneity and SHAP global average absolute value quantitative variables, strictly distinguish between model fitting weight and long-term factor pricing premium, and the two main lines of analysis should not be confused with each other.

4. Empirical Results and Analysis

4.1 Descriptive Statistics of the Main Variables

The effective observation of the whole sample is a total of 117,540 groups, with the average return of individual shares in the next month -0.0067, and the standard deviation is 0.1313, which visually reflects the high noise characteristics of the monthly return of A-shares; The deviation of the original extreme value of the monthly change rate and the monthly momentum is extremely large. This article standardizes preprocessing all core variables: first, the upper and lower limits of the variables with extreme change, such as monthly return, change of hands, momentum, etc., are cut off, and the extreme error values that do not conform to the realistic logic of the A-share market are eliminated; Subsequently, 1% and 99% percential bilateral shrinkage treatment is carried out on all variables to further weaken the interference of extreme observation on model estimation and ensure the robustness of empirical results.

Table 1: Descriptive statistics of main variables

Variable	Observations	Mean	Standard Deviation	Minimum	Median	Maximum
Next-month stock return	117540	-0.0066	0.1311	-0.2987	-0.0171	0.5843
Log market capitalization	117540	15.7776	1.1144	12.3869	15.6437	18.9772
Book-to-market ratio	117540	0.4154	0.2205	0.0799	0.4008	20.5246
Return on assets (ROA)	117540	0.0574	0.1338	-0.1320	0.0462	0.1408
Monthly turnover rate	117540	0.0015	0.0194	-0.0065	0.0003	1.0000
Monthly momentum	117540	0.0082	0.0636	-0.0244	-0.0006	1.0000

Note: This table reports the descriptive statistics of the variables after outlier treatment and two-sided winsorization at the 1st and 99th percentiles. All variables are reported in decimal form, and their distributions are broadly consistent with the typical characteristics of the China A-share market.

4.2 Preliminary Test for Multicollinearity among Factors

This study employs the Pearson correlation coefficient and the Variance Inflation Factor (VIF) to assess multicollinearity among the variables; the maximum pairwise correlation coefficient between factors is only 0.1928, and the VIF values for all factors are below 10. Only the logarithm of market capitalization and the book-to-market ratio exhibit mild multicollinearity; no multicollinearity is present among the other variables, allowing them to be directly incorporated into the three types of models for regression fitting.

Table 2: Pearson Correlation Coefficient Matrix and VIF Collinearity Test

Factor	Book-to-Market Ratio	Return on Assets (ROA)	Monthly Turnover Rate	Log Market Capitalization	Monthly Momentum	VIF	Collinearity Assessment
Book-to-market ratio	1.0000	-0.1262	0.0361	0.1928	0.0068	4.8112	Mild multicollinearity
Return on assets (ROA)	-0.1262	1.0000	0.0444	0.0745	0.0260	1.2160	No significant multicollinearity
Monthly turnover rate	0.0361	0.0444	1.0000	0.0045	0.1723	1.0404	No significant multicollinearity
Log market capitalization	0.1928	0.0745	0.0045	1.0000	-0.0061	5.2254	Mild multicollinearity
Monthly momentum	0.0068	0.0260	0.1723	-0.0061	1.0000	1.0480	No significant multicollinearity

Notes: (a) The first five columns report the pairwise Pearson correlation coefficients among the factors. (b) Following the commonly adopted criterion, a VIF value below 10 indicates the absence of severe multicollinearity. (c) All statistics were computed in batches using the statsmodels package in Python.

4.3 Monthly Coefficient Estimates of the Linear Models

The LnSize coefficient in the FM model is significantly negative, consistent with established domestic and international findings regarding the size premium [12]; ROA exhibits the strongest linear predictive power. The unconstrained FM lacks a regularization term and is prone to fitting noise; while Lasso significantly shrinks factor coefficients, their signs remain unchanged. This indicates that regularization suppresses random disturbances without altering the direction of factor return predictions, thereby reducing noise and enhancing generalization capability.

Table 3: Average monthly regression coefficients

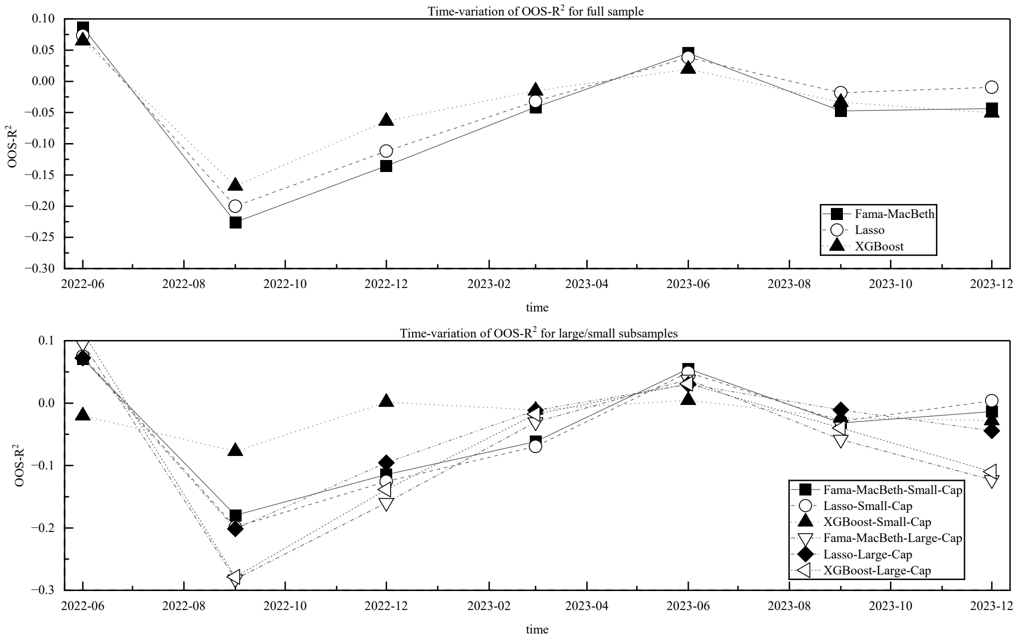
Factor	Average FM Coefficient	Average Lasso–FM Coefficient
Log market capitalization	-0.000265	-0.000108
Book-to-market ratio	0.006044	0.001740
Return on assets (ROA)	0.019875	0.000840
Monthly turnover rate	0.000719	0.000438
Monthly momentum	0.000055	0.000059

4.4 Multidimensional Prediction Results under the Benchmark Window and Analysis of Market Capitalization Heterogeneity

4.4.1 Overall Predictive Performance under Different Rolling Window Lengths

As shown in Figure 1, under the 60-month benchmark window, the average values of OOS-R^2 of FM, Lasso-FM and XGBoost are -0.0241, -0.0097 and -0.0328 respectively. The out-of-sample decision coefficients of the three types of models are all negative, indicating that the absolute level prediction ability of the monthly return of individual stocks by relying on the five basic factors of this group is limited. Although the OOS-R^2 mean of Lasso-FM is relatively higher, the Newey-West robust t-test shows that there is no statistically significant difference between the models, and it cannot be determined that Lasso-FM has a deterministic advantage. At the same time, the indicators are differentiated: XGBoost’s IC, Rank IC, the annualized Sharpe ratio of the decile long–short portfolio and other sorting and Portfolio performance indicators are better than Lasso-FM.

Figure 1: Time series comparison of model OOS-R² under different rolling windows



Note: The 36-month and 48-month windows are used only for parameter sensitivity testing, not for robustness checks.

The differentiation of the results of the two types of indicators comes from the different evaluation logic. OOS-R² measures the absolute prediction error of returns, and the prediction deviation of individual samples will lower the index; IC and Rank IC only pay attention to the relative sorting of individual stock earnings, and do not require accurate prediction of the return magnitude. Even if the absolute fitting effect is poor, as long as the model can distinguish the relative level of stock returns, it can still get positive IC and Rank IC, and obtain better risk-adjusted returns through long–short portfolios, which also explains the phenomenon that a higher Sharpe ratio can still occur when OOS-R² is negative.

The results of this paper are not exactly the same as the conclusion of XGBoost comprehensive dominant obtained by most high-dimensional factor studies. Under the sample setting of the five basic factors in this article, the high-order interaction information that can be mined by the tree model is limited, the monthly return noise of superimposed A-shares is high, and the XGBoost with a high degree of freedom is more likely to fit the random disturbance in the training set, resulting in the weakening of the absolute fit performance outside the sample. It should be noted that the above is only a potential mechanism under this sample. There is no universal and fixed threshold for the number of factors. If the factor pool or sample interval is replaced, the relative performance of the model may change.

Compared with existing studies, Gu Songnan can give full play to the ability to capture high-order interactions of XGBoost based on high-dimensional derived factors, and get the conclusion that the tree model has a better prediction effect [2]; Li Bin carried out Lasso modeling based on low-dimensional factors, which confirmed that the regular linear model has good generalization performance, which is mutually confirmed with the observations in this article [4]. Jiang Fuwei and others also pointed out that the dimension of the factor pool is the key premise for determining the relative performance of the machine learning model. The conclusion obtained by the high dimension cannot be directly copied to the modeling scenario of the low-dimensional basic factors of small and medium-sized institutions [5].

It can be seen that there is no absolute dominant model in low-dimensional scenes, and a certain type of model cannot be directly prioritized. Judging from the numerical results of the sample, Lasso-FM has a relatively higher OOS-R² average, but the difference did not pass the Newey-West statistical significance test; XGBoost reflects a better cross-section stock selection ranking performance. In practice, it should be selected in combination with investment goals, computing power conditions and the size of the factor pool; When a large-scale derivative factor library can be built in the future, the high-order interactive mining ability of XGBoost can be fully released.

Based on the above results, the three groups of key issues can be further sorted out. First, the differentiated results of this article and most high-dimensional factor studies do not mean that the XGBoost model itself is invalid, but is constrained by factor input conditions: The high-dimensional derivative factor pool can provide sufficient high-order interaction information for the tree model, which can give full play to its nonlinear mining advantages; However, under the low-dimensional condition of only five basic factors in this article, the interactive information that can be mined is limited. In the high noise environment of the monthly return of A-shares, it is easier for the high-degree-of-freedom model to capture random disturbances in the training set, which eventually causes differences in the conclusions of the two types of literature. This phenomenon also shows that the conclusion of the machine learning model is highly dependent on the factor set setting. Second, the results have practical implications for the modeling of small and medium-sized institutions: There is no universal optimal model. If the research target focuses on the relative ranking of stocks and builds a long–short portfolio, you can pay attention to XGBoost; If you want to focus more on the fitting performance of the absolute level of income, you can refer to Lasso-FM; Model selection needs to match the computing power of the institution, factor resources and investment targets, and cannot simply copy the conclusions of high-dimensional literature. Third, the characteristics of the A-share market are an important real soil that causes indicator divergence and model performance differentiation. The market share of retail investors is relatively high, and the theme speculation brings a lot of short-term noise. The monthly real return signal of individual stocks is covered up by noise, so it is difficult for the model to accurately predict the absolute size of the return (OOS-R^2 is negative); However, the noise is more reflected in the specific disturbance of individual stocks, which does not necessarily completely destroy the order of relative strength and weakness between stocks. Therefore, even if the absolute fit fails, it can still get a positive IC and a higher multi-air Sharp. The above are all empirical inferences obtained under this sample. If the factor set or sample interval is replaced, the relative performance of the model may change.

The proportion of retail transactions in the A-share market is high, and the theme speculation brings a lot of return noise [1,9]. Lasso compresses the false coefficients driven by noise through the L1 regularization; while XGBoost has a stronger fitting ability and is easier to capture pseudo-correlations in a high-noise environment. This inference is only a potential path under this sample. The institutional tracking of the large-market target is sufficient, and the stock price is less affected by speculation; The problem of information asymmetry of small-cap stocks is prominent, and effective signals are easily diluted by capital noise, which provides a preliminary understanding perspective for the subsequent market value grouping heterogeneity results of this chapter.

4.4.2 Predictive Performance across Market Capitalization Groups

As shown in Table 4, the large-cap stocks OOS-R^2, IC and Rank IC are superior to the small-cap stocks in an all-round way. The IC and Rank IC of the small-cap stocks FM turn into negative values, and the corresponding theme hype brings short-term profit reversal. Combined with the theory of information economics, the disclosure of large-market information is complete, the institution is continuously tracked, and the fundamental signal is weakly impacted by speculation; Small-sized retail investors dominate, information lags behind, and effective profit signals are diluted by random funds. This article does not have institutional shareholdings, and analysts cover agent variables. It is only a correlation observation and no causal judgment. The small-cap stocks XGBoost OOS-R^2 fell to -0.0422, and the overfitting defect was further amplified in the noisy environment. At the practical level, setting up circulating market value filtering can reduce the pullback; improving the rules of small-cap information disclosure on the regulatory side can narrow the stratification gap of market information.

Table 4: Average predictive indicators and market cap heterogeneity (60-month window)

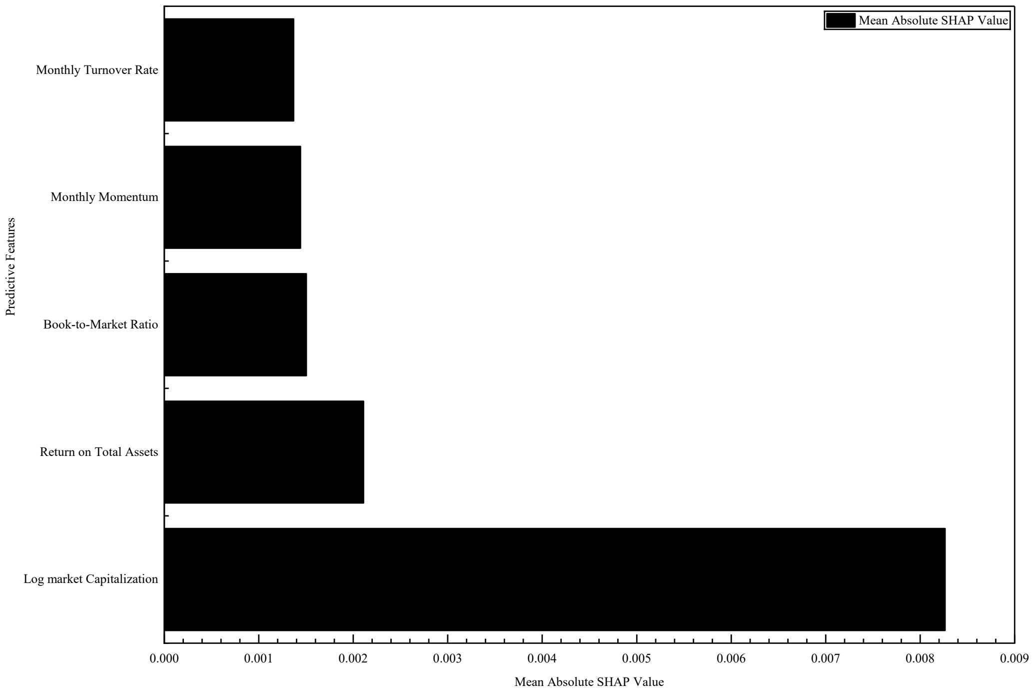
Sample Group	Performance Metric	Traditional FM	Lasso-FM	XGBoost
Full sample	Mean OOS-R^2	-0.0241	-0.0097	-0.0328
Full sample	Mean IC	0.0304	0.0099	0.0503
Full sample	Mean Rank IC	0.0366	0.0098	0.0456
Full sample	Annualized Sharpe ratio	0.9432	0.0662	1.3679
Large-cap stocks	OOS-R^2	-0.0329	-0.0037	-0.0307
Large-cap stocks	Mean IC	0.0365	0.0117	0.0316
Large-cap stocks	Rank IC	0.0455	0.0135	0.0263
Large-cap stocks	Annualized Sharpe ratio	0.6991	1.1607	1.8739
Small-cap stocks	OOS-R^2	-0.0172	-0.0176	-0.0422

Sample Group	Performance Metric	Traditional FM	Lasso-FM	XGBoost
Small-cap stocks	Mean IC	-0.0048	-0.0029	0.0204
Small-cap stocks	Rank IC	-0.0166	0.0105	0.0093
Small-cap stocks	Annualized Sharpe ratio	-0.5743	0.9162	0.3645

Notes: OOS-R2R^2 denotes the out-of-sample coefficient of determination. IC and Rank IC are reported as monthly averages. The annualized Sharpe ratio is calculated based on decile long–short portfolios and does not account for transaction costs. The results are generated from rolling-window forecasts implemented in Python and are reported to four decimal places.

4.4.3 Analysis of Feature Contributions in the XGBoost Model

Figure 2: Global SHAP contribution of important predictive features in XGBoost model



Note: The x-axis represents the mean absolute SHAP value, and the y-axis represents the five major fundamental factors; the bar length indicates the magnitude of the variable's contribution to the overall model's predictions.

Combined with the SHAP decomposition results in Figure 2, the SHAP mean is sorted: log market capitalization > ROA > BM > monthly momentum > monthly turnover rate. Explained from the theory of information stratification: the circulating market value directly distinguishes the information transparency and noise hierarchy of the target, so the prediction weight in the model is the highest; ROA reflects long-term operation, but short-term stock prices are susceptible to subject disturbances, and the contribution is secondary. Strictly distinguish the concept: SHAP only measures the internal fitting contribution of the model, which does not represent the long-term asset pricing premium. If you need to test the pricing ability of the factor, you need to build a long-term layered portfolio separately, and the high SHAP variable cannot be defined as the core pricing factor.

4.5 Sensitivity Analysis of Rolling Window Length and Robustness Checks

Replace the 36-month window to do the parameter sensitivity test, and intercept the 2018-2023 segment samples to carry out the robustness test. The results of the window sensitivity test and the segment robustness test show that the indicators under different settings remain the same, but there is no statistically significant gap between the advantages and disadvantages between various models. The overall forecasting index is superior to that of the small-cap stocks. The observation is robust. All five types of independent robustness tests have been passed, and the conclusions are reliable. The results are shown in Table 5.

Table 5: Parameter Sensitivity Test and Subsample Robustness Test Results

Sample Type	Sample Group	Performance Metric	Traditional FM	Lasso-FM	XGBoost
36-month rolling window (Sensitivity analysis)	Full sample	OOS-R^2	-0.0517	-0.0372	-0.0351
36-month rolling window (Sensitivity analysis)	Full sample	Mean IC	0.0339	0.0004	0.0602
36-month rolling window (Sensitivity analysis)	Full sample	Rank IC	0.0613	-0.0101	0.0596
36-month rolling window (Sensitivity analysis)	Full sample	Annualized Sharpe ratio	0.7569	-2.9241	1.5419
36-month rolling window (Sensitivity analysis)	Large-cap stocks	OOS-R^2	-0.0743	-0.0374	-0.0628
36-month rolling window (Sensitivity analysis)	Large-cap stocks	Mean IC	0.0322	0.0098	0.0457
36-month rolling window (Sensitivity analysis)	Large-cap stocks	Rank IC	0.0585	-0.0179	0.0369
36-month rolling window (Sensitivity analysis)	Large-cap stocks	Annualized Sharpe ratio	1.3304	-1.6556	2.0728
36-month rolling window (Sensitivity analysis)	Small-cap stocks	OOS-R^2	-0.0394	-0.0422	-0.0220
36-month rolling window (Sensitivity analysis)	Small-cap stocks	Mean IC	0.0208	-0.0044	0.0372
36-month rolling window (Sensitivity analysis)	Small-cap stocks	Rank IC	0.0122	-0.0550	0.0257
Subsample robustness check (2018–2023)	Full sample	OOS-R^2	-0.0247	-0.0104	-0.0352
Subsample robustness check (2018–2023)	Full sample	Rank IC	0.0218	0.0113	0.0250
Subsample robustness check (2018–2023)	Full sample	Annualized Sharpe ratio	0.5271	0.2664	1.1758

Notes: *The 36-month and 48-month rolling windows are used solely for hyperparameter sensitivity analysis and are not treated as independent robustness checks.*

5. Conclusions and Implications

5.1 Main Findings

1.Under the setting of the monthly five basic low-dimensional factors of A-shares from 2013 to 2023, the average value of OOS-R^2 outside the samples of FM, Lasso-FM and XGBoost samples is all less than 0, and the model's absolute numerical interpretation of the monthly return of individual stocks is weak; However, there is a significant differentiation in the two types of evaluation systems: the average value of the OOS-R^2 sample of Lasso-FM is relatively higher, but the difference is not statistically significant; XGBoost's IC, Rank IC and long–short annualization Sharp ratio are in the overall lead, and the stock selection sorting and portfolio return performance are better. The difference between the two is rooted in the different index evaluation logic: OOS-R^2 measures the degree of accurate fit of the yield, which is extremely vulnerable to the random noise of the A-share market; IC and Rank IC only distinguish the relative strength of stock returns, and there is no need to accurately predict the real level of returns, thus forming a typical departure of "negative OOS-R^2 with positive excess combination returns". The Newey-West test shows that there is no statistical level of significant superiority or inferiority in the comprehensive performance of the three types of models, and the model needs to be selected in combination with the investment target.

2.The model predicts that there is significant market value heterogeneity, and the performance of all evaluation indicators of large-cap stocks is better than that of small-cap stocks. The information disclosure of the large-market target is perfect, the institutional tracking is sufficient, and the fundamental forecast signal is less interfered by short-term speculation; Small-market is mainly based on retail trading and thematic speculation. Information asymmetry dilutes effective pricing signals, and the problem of XGBoost's over-fitting in small-market samples is further aggravated. SHAP feature decomposition shows that market capitalization and ROA are the core prediction variables in the model, which only represent the model fitting weight, which is not equivalent to the long-term factor risk premium.

3.The dimension of the factor pool is the core boundary condition that determines the relative performance of the linearity and tree model. Under the sample setting of the five basic factors in this article, there is very little interaction between the high-order factors that can be mined by XGBoost, superimposed on the high noise characteristics of A-share monthly return, and the random disturbance of the high-freedom tree model easy memory training set, and the absolute fit performance outside the sample is not as good as the Lasso regular linear model. This law is only applicable to low-dimensional basic factor scenarios, and cannot be generalized to the hundred-dimensional derivative factor environment, which explains the intrinsic reasons for the differences between this article and the conclusions of most high-dimensional machine learning stock selection literature; The results of window sensitivity and five types of independent robustness tests show that the above empirical observations obtained from this sample are robust.

The incremental value of this article is mainly reflected in three aspects: at the theoretical level, it clarifies the inherent logical differences between the two types of evaluation indicators. Under the framework of unified rolling and lag backtesting, it is confirmed that the three types of models OOS-R^2 are all negative, but the XGBoost sorting index is better than that of multi-empty Sharp ratio, which explains the contradiction that "negative OOS-R^2 can still obtain excess returns" and enriches the interpretation of multiple prediction and evaluation indicators in the existing literature; At the empirical boundary level, based on sample observations, it is revealed that the factor dimension is an important constraint that affects the relative performance of linear and tree models, which provides differentiated empirical evidence for the inertial cognition that "machine learning is naturally superior to linear models". At the practical level, it provides a targeted lightweight modeling scheme for small and medium-sized institutions, and gives model selection ideas according to the two types of goals of absolute return fitting, stock selection ranking and excess return; At the same time, a standardized sliding rolling backtest process to avoid information leakage is built to provide a reproducible research paradigm for quantitative practices with limited computing power and factor resources.

5.2 Practical Implications

a) Quantitative practical level: There is obvious index differentiation in empirical evidence: XGBoost stock selection sorting, long–short Sharp performance is stronger, and the OOS-R^2 sample average of Lasso-FM is relatively higher, but the difference is not statistically significant. There is no global optimal model, and it needs to be selected in combination with the target and computing power:

1. If you aim to select long–short stocks and obtain relative excess returns, you can choose XGBoost; in the large-market stock sample, the long–short Sharp advantage of this model is particularly prominent.

2. If the computing power resources are limited, only the basic factor can be used, and the absolute fit stability of the gain and the overfitting risk control can be more important, you can choose the lightweight Lasso-FM model, relying on the L1 regularization rule to suppress the false coefficient brought by noise, which can reduce the predicted fluctuation in a small-scale high-noise environment.

3. Uniformly adopt circulating market value filtration to reduce the combination withdrawal; after expanding the derived high-dimensional factor library, switch XGBoost to give full play to the nonlinear interactive mining ability.

b) At the regulatory level: insufficient information disclosure of small-chap stocks and loud speculation noise of retail investors widen the performance gap of small-scale stocks. Supervision can improve the normalized information disclosure system of small-markets, reduce information asymmetry, alleviate the disturbance of short-term speculation on stock prices, and improve the effectiveness of small-market factor prediction.

5.3 Limitations and Directions for Future Research

This paper only uses 5 basic factors to carry out low-dimensional modeling, and mainstream machine learning-related research is mostly based on high-dimensional factor data sets. The parallel reference literature under similar low-dimensional conditions is relatively limited. The marginal contribution of this article has been explained above, and the research limitations are sorted out as follows: (1) Without the construction of a high-dimensional derived factor library, it is difficult to give full play to Lasso's variable screening effect and XGBoost's ability to mine the high-order interaction of factors. (2) The hyperparameter reference of the machine learning model has been researched to set a fixed value, and the grid search parameter adjustment is not carried out based on the timing verification set. (3) The lack of intermediary agent variables such as institutional shareholding and analyst coverage makes it impossible to quantify the transmission path behind the heterogeneity of market value. (4) Sample testing is carried out without distinguishing between bull market, bear market and shock market. ⑤The two sets of independent test frameworks of short-term income forecasting and long-term factor pricing are not distinguished.

In the future, we can construct a high-dimensional factor data set, introduce an independent timing verification set to carry out hyperparameter optimization, supplement the investor's structural agent variables, distinguish the performance of different market comparison models, and further improve the layered empirical research of linear and tree models.

References

[1] Wu, S. N., & Xu, N. H. (2004). A comparative study of rational and irrational asset pricing models: Evidence from China's stock market. *Economic Research Journal*, *39*(6), 105–116.

[2] Gu, S. N. (2023). Application of machine learning to cross-sectional return prediction in China's A-share market. *Statistics & Decision*, *39*(12), 156–161.

[3] Qiao, Z., Zhuo, R. S., Ge, Y., & Liu, Y. S. (2025). Corporate fundamental valuation mismatch anomaly based on machine learning. *Systems Engineering—Theory & Practice*, *45*(9), 2912–2933.

[4] Li, B., Shao, X. Y., & Li, Y. Y. (2019). Machine learning–driven fundamental quantitative investment. *China Industrial Economics*, *(8)*, 61–79.

[5] Jiang, F. W., Xue, H., & Zhou, M. (2022). Does big data improve the pricing performance of multi-factor models? Evidence from China's A-share market based on machine learning. *Systems Engineering—Theory & Practice*, *42*(8), 2037–2048.

[6] Bryzgalova, S., Pelger, M., & Zhu, J. (2025). Forest through the trees: Building cross-sections of stock returns. *The Journal of Finance*, *80*(5), 2447–2506.

[7] Wang, X., Wu, F., Huang, D. J., Chen, Y., & Xiang, Z. (2022). A time-varying multi-factor model for China's A-share market. *Journal of Southwest University (Social Sciences Edition)*, *48*(6), 137–152.

[8] Tang, G. H., Zhu, L., Liao, C. F., & Jiang, F. W. (2024). Asset pricing based on autoencoder machine learning: A financial big data perspective on China's stock market. *Journal of Management Sciences in China*, *27*(9), 82–97.

[9] Liu, X. D., Liu, C., Liu, S. C., & Lu, J. J. (2014). Does herd behavior exacerbate stock price volatility? *Systems Engineering—Theory & Practice*, *34*(6), 1361–1368.

[10] Su, Z., Lu, M., & Li, D. X. (2017). Empirical applications of deep learning in finance: Developments, contributions, and prospects. *Journal of Financial Research*, *(5)*, 111–126.

[11] Ma, T., Jiang, F. W., & Tang, G. H. (2022). Deep learning and factor investing in China's stock market: Based on generative adversarial networks. *China Economic Quarterly*, *22*(3), 819–842.

[12] Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427–465.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.